# A literature review of association rules in mining

Technology, Computer

# Abstract

Mining association rules is an essential job for information discovery. Past transaction data can be analyzed to discover client behaviors such that the superiority of business decision can be improved. The approach of mining association rules focuses on discovering large item sets, which are groups of items that come into view together in a sufficient number of dealings. Association rules are if/then statements that help uncover relationships between seemingly unrelated data in a information repository. In this paper we will show by experimental results the behavior of apriori algorithm. Weshall describes the basic concepts of association rules mining, the basic model of mining association rules. Finally, this paper describes the association rules mining and its techniques.


Introduction

Association rules mining is an important task in data mining. It is a popular and well researched method for discovering strong associations between variables in large databases. It is intended to discover strong rules between different variables in databases. A large amount of data can easily be analyzed to discover customer purchasing behavior which improve business behavior. The goal of the association rules mining is to identify items that are bought together by sufficiently many customers. The strong relation between different items in the market are existing like “ the peoples who buy milk also tends to buy bread and eggs” in this sentence there is a relation between milk and bread. So association rules are used to identify these relationships between items for the improvement of business behavior.

Association rules can be expressed as: R: X==> Y, where: X âŠ, I, Y I, and

X, it is said that if the item set X occurs in a transaction, then Y will inevitably appears in the transaction. Therefore, X is called a prerequisite for the rule; Y is the result of the rule. Support and Confidence are two different interestingness for measure. Support of an item I is the number of transactions that support (contains) I, and Confidence compares the number of times the pair was purchased to the number of times one of the items in the pair was purchased.

In probability terms this is referred to as the conditional probability of the pair. For example, if a supermarket database has 100, 000 point-of-sale transactions out of which 2, 000 include both items A and B and 800 of these include item C, the association rule " If A and B are purchased then the item C is also purchased on the same trip" has a support of 800 transactions (alternatively 0. 8% = 800/100, 000) and a confidence of 40% (= 800/2, 000).

A huge number of association rules can be identified if the database is large. So for minimizing association rules minimum Support and Confidence are considered, both are specified by the user which help us to and valuable rules from database.

Association Rule Mining Algorithms

A-priori Algorithm

Principle of Apriori Algorithm: " If an item set is frequent, then all of its subsets must also be frequent".

Apriori algorithm is a classical and breadth first search association rules algorithm. This algorithm was first proposed by Agrawal et al in 1993. Apriori algorithm strategy is to separate association rule mining tasks into two steps:

First discover frequent item sets, and the second is the Generating of Association rules, it extracts high confidence rules from the frequent item sets. The first step for mining frequent item sets the algorithm will produce a large number of Items; the algorithm will execute K iterations where K is the number of items in the second iteration the algorithm produce some frequent item sets with the first selected frequent item set. After the K iteration the algorithm produce the superset of all frequent items.

Here the basic idea of generating frequent item sets is: First step, statistics the frequency of the set with an element, and identify those item sets that is not less than the minimum support, that is, the maximum one-dimensional item sets. Then start the cycle processing from the second step until no more maximum item sets generated. The cycle is: in the first step k, k-dimensional candidate is generated form (k-1) dimensional maximum item sets, and then scans the database

to get the candidate item set support, and compare with the minimum support, k-dimensional maximum set is found. The apriori algorithm takes advantage of the fact that any subset of a frequent item set is also a frequent item set, therefore it reduce the number of candidates being considered by only exploring the item sets whose support count is greater than the minimum support count snf all infrequent item set can be pruned if

it has an infrequent subsets. Apriorialgorithm is breadth-first algorithm, therefore if the database is too large then it suffer from a number of inefficiencies by creating a large number of subsets.

Apriori algorithm uses sets intersections to determine support values. It determines the support values of all (K-1) candidates before counting the K candidates. The dataset may be too large thus the problem is that the resulted frequent item set may be exceeded with main memory and wasting of time to

Figure 1: Apriori algorithm Pseudo code.

Figure 2: Improved version of Apriori Algorithm

Hold a large number of candidate set with much frequent item sets. So to overcome this problem the dataset is partitioned in different chunks and each chunk is treated independently. And then the resulted frequent items are merged with one extra scan.

Applications: Basket data analysis, cross-marketing, catalog design, sale campaign analysis, Web log (click stream) analysis, and DNA sequence analysis.

2. 1. 1. Variation in Apriori Algorithm

The limitation of appriori algorithm is improved by the improved version ofappriori algorithm. It is to be defined as: Suppose Ck is the candidate item set of size k, and Lk is the frequent item set of size k in the proposed approach the algorithm is improved by reduce the time consuming for

candidates item set generation. Here the algorithm firstly scan all transactions to generate L1 which contains all items, and found their support and transaction ID, and then L1 is used as a helper to generate L2, L3…. Lk, then generate C2 by joining L1 * L1 to construct 2-itemset C(x, y) where x, y are the items of C2. Before scanning all transaction records to count the support count of each candidate, use L1 to get the transaction IDs of the minimum support count between x and y, and thus scan for C2 only in these specific transactions. The same thing for C3, construct 3-itemset C (x, y, z), where x, y and z are the items of C3 and use L1 to get the transaction IDs of the minimum support count between x, y and z, then scan for C3 only in these specific transactions and repeat these steps until no new frequent item sets are identified.

FP-Growth Algorithm

FP-growth algorithm is one of the latest and most efficient algorithms in depth-first algorithm. It allows frequent item sets discovery without candidates item set generation. It is a two step approaches, first build a compact data structure called FP-tree and then extract the frequent item sets directly from the FP-tree.

Compared with Apriori Algorithm, FP-growth has the following advantages:

To avoid multiple dataset scanning it scan only the dataset twice. It increases space and time efficiency. But its difficulty lies in large and sparse datasets, in the mining processing and recursive computations require considerable space.

Applications: Basket data analysis, frequent patterns.

Association Rules Applications

3. 1. Market Based Data Analysis

A typical and widely-used example of association rule mining is market basket analysis. It is a technique that discovers relationships between pairs of products purchased together. The technique can be used to identify the items having strong relationship. The idea behind market basket analysis is simple, simply examine the order of products have been purchased together. For example in market basket analysis the fact might be uncover that if “ a customer buy milkalso tends to buy breads”. So using this information we might organize our store that milk and bread next to each other.

For doing market basket analysis there is some couple of measures is used, which is frequency, minimum Support and minimum Confidence, frequency is the number of times two products were purchased together, and minimum Support and Confidence are discussed before. Market Basket Analysis print report about given items, for example if we need to find relationship of Milk with others Bread, Eggs and Cheeses, then market basket analysis print a reports. TheReport consists of the products name, Frequency, Support and Confidence.

Market Basket Analysis: Milk

Product

Frequency

Support

Confidence

Breeds

820

82%

91. 1%

Cheese

800

80%

23. 5%

Eggs

750

75%

34%

The higher the confidence means that there is a probably of strong relationship between the products. In the above example confidence of Milk and Breads shows that in 91% of transactions Milk and Bread are sold together.

3. 2 Customer Relationship Management (CRM)

Customer Relationship Management is a combination of business process and technology that seeks to understand a company’s customers from the perspective of who they are, what they do, and what they are like. Here we are focusing on the CRM of banking sector, which are focused to find the preference of different customer, to provide services to the customer to enhance cohision between customers and the bank. Association rules are used to identify customer preferences and customer behavior.

Conclusion

Association rules mining are a popular and well researched method for discovering strong associations between variables in large databases. In this paper we have describe Association rules mining which is the important task of data mining. Then we describes techniques for the Association rules which is apriori and FP-growth algorithm, the limitation of apriori algorithm was inefficiency in case of large database by checking all k-1 items, the improve apriorialgorithm overcome this problem by finding transactions ids of every frequent item and then for k+1 set generation used that k-1 table and search only those transactions in which the current item are exist. According to this the efficiency of the algorithm is improved. Then we discuss FP-growth algorithm which is depth first search and fastest algorithm, it uses tree to find frequent item sets. In the last the application of association rules Market basket data analysis and Customer Relationship management are discussed in this paper.

References

Mohammed Al-Maolegi, Bassam Arkok Jordon, “ An improved apriori algorithm for association rules” International Journal on Natural Language Computing (IJNLC) Vol. 3, No. 1, February 2014.

Ruowu Zhong and Huiping Wang China “ Research of Commonly Used Association Rules Mining Algorithm in Data Mining 2012.

S. Rao, R. Gupta, “ Implementing Improved Algorithm Over APRIORI Data Mining Association Rule Algorithm” International Journal of Computer Science And Technology, pp. 489-493, Mar. 2012.

Jiawei Lian , Michelins Kamber. Data mining: Concepts and Techniques [M]. America: Morgan Kaufman Publishers, 2000.

Market basket data analysis “”

Show-Jane Yen and Arbee L. P. Chen Taiwan “ An Efficient Data Mining Technique for Discovering Interesting Association Rules” 2010